Final Project

Christophe

November 2, 2015

```r
# load data
library(RCurl)
url <- paste(url,"STD_Rate_Data_File.csv", sep = '')
std.data <- getURL(url)
std.data <- read.csv(textConnection(std.data), stringsAsFactors = FALSE)
df <- data.frame(std.data)
df <- subset(df, df$Race.Ethnicity != 'Unknown'
             & df$Gender.Code != 'U' & !is.na(df$STD.Cases)
             & df$STD.Cases != "Suppressed")
df <- df[!is.na(df$STD.Cases),]
df$STD.Cases <- as.numeric(df$STD.Cases)
df$Rate <- as.numeric(df$Rate)
```

Part 1 - Introduction:

Chlamydia is a common sexually transmitted infection ("STI") that can infect both men and women. The infection is directly responsibile for negative health outcomes when left untreated. The infection can damage a woman's reproductive system by making it near impossible for her to get pregnant later in life, and it can cause a potentially fatal ectopic pregnancy (pregnancy that occurs outside the womb). [1]

Due to these negative health outcomes, Chlamydia can cause a real economic hardship for a community. In 2000, the direct care cost for Chlamydia for just American Youth (ages 15 - 24) was $248.4 million [2]. This sum is a staggering cost to the public that requires attention from public health officials. Especially, since Chlamydia is also easily diagnosed and treated but remains a substantial burden in the US [3].

In this data project, we will examine the correlation between ethnicity and Chlamydia rates in the United States. It is my hope that by determining a correlation we may be able to identify hot spots of infection for further analysis and possibly assist in informing public health decisions in reducing new incidence of Chlamydia.

Part 2 - Data:

Data Collection

Beginning in 2003, all 50 states and the District of Columbia had converted from summary hardcopy reporting to electronic case-specific reporting via NETSS (with the exception of congenital syphilis, which is still reported on hardcopy forms by several areas). Also, before 1996, Chlamydia reporting was voluntary, and thus sporadic. From 1995 - 2000, upstate New York did not report Chlamydia, and thus the total US denominator population for Chlamydia excludes the New York state population for the years 1995 - 2000. However, New York City did report Chlamydia before the year 2000, and the New York City population and Chlamydia cases are included for 1984 - 2000. (Sexually Transmitted Disease Morbidity by Race and Age. (n.d.). Retrieved October 17, 2015.)

It is unclear how many layers of abstraction exists for the data set. It seems that the information is compiled from provider level reports of each patient and then aggregated by the various health agencies and reported to CDC. However, it could be the case that each patient or incidence is reported to the CDC and the CDC does the aggregation. This difference in reporting would only affect our confidence about the total accuracy of the information.

Cases

The number of cases and disease incidence rates are reported by year, gender of patient, race/ethnicity, type of STD, and state.

Our investigation will focus on only the infection of Chlamydia in the year 2013 as shown in the below table.

```
cases_table <- aggregate(list('Cases in 2013' = df$STD.Cases),
                 by = list(Ethnicity = df$Race.Ethnicity), FUN=sum)
cases_table <- cases_table[order(cases_table$Ethnicity),]
kable(cases_table, caption = "Number of Cases Reported in 2013 by Ethnicity")
```

Table 1: Number of Cases Reported in 2013 by Ethnicity

Ethnicity	Cases.in.2013
American Indian or Alaska Native	17407
Asian or Pacific Islander	20415
Black or African American	431788
Hispanic	197550
White	346140

The minimum number of cases reported for an ethnicity is 17,407 and the maximum reported cases by an ethnicity is 431,788. This gives us the range of 414,381 which is a very large range of cases between enthnicity. Therefore, for our data analysis we will need to normalize the data for the population differences and look at `rates` of infection.

Variables:

We will be examining the `state` variable which is categorical and provides a geographical observation. Each `state` had different laws and public health policies, finding a high correlation due to state for rates may provide insight to which `state` may be further examined for its handling of Chlymadia. Note that the District of Columbia is included which will provide us with 51 `state` variables whereas the US technically only has 50 states. In the analysis we further group the states into regions.

```
cases_by_state <- aggregate(list('Cases in 2013' = df$STD.Cases),
                            by = list(State = df$State), FUN=sum)
kable(cases_by_state, caption = "Cases by State")
```

Table 2: Cases by State

State	Cases.in.2013
Alabama	19745
Alaska	5725
Arizona	24160
Arkansas	13291
California	106240
Colorado	11724
Connecticut	4639
Delaware	4969
District of Columbia	3544
Florida	60372

State	Cases.in.2013
Georgia	29335
Hawaii	2842
Idaho	3376
Illinois	50739
Indiana	25008
Iowa	9899
Kansas	6453
Kentucky	11785
Louisiana	28127
Maine	2433
Maryland	19819
Massachusetts	14047
Michigan	29494
Minnesota	12930
Mississippi	13906
Missouri	22780
Montana	3563
Nebraska	4548
Nevada	7702
New Hampshire	2745
New Jersey	13688
New Mexico	7707
New York	53958
North Carolina	40371
North Dakota	2280
Ohio	37082
Oklahoma	14962
Oregon	10756
Pennsylvania	32769
Rhode Island	3503
South Carolina	17283
South Dakota	3336
Tennessee	29499
Texas	115080
Utah	7504
Vermont	1803
Virginia	23070
Washington	18308
West Virginia	4354
Wisconsin	18352
Wyoming	1695

The second variable we will be analyzing is `Race.Ethnicity` which is also categorical. `Race.Ethnicity` may provide insight to health outcome disparities among certain ethnic groups.

```
kable(unique(df$Race.Ethnicity), col.names = "Ethnicity", caption = "Ethnicities Variables" )
```

Table 3: Ethnicities Variables

Ethnicity
American Indian or Alaska Native
Asian or Pacific Islander
Black or African American
Hispanic
White

These two variables will be utilized to determine if there is difference between the rates of the observations for these categorical variables.

Type of study

This study is an observational study on Sexually Transmitted Disease (STD) morbidity case records reported to the National Center for HIV/AIDS, Viral Hepatitis, STD, and TB Prevention (NCHHSTP), Centers for Disease Control and Prevention (CDC).

We are able to determine that this is an observational study because the cases are self-reported and we have not controlled for any confounding factors. Since the sampling was through self-reporting and there is no comparison to a control group, we will be limited in only being able to infer correlation and not causation.

Data Source

The data source is publicly available from the US Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for HIV, STD and TB Prevention (NCHSTP), Division of STD/HIV Prevention, Sexually Transmitted Disease Morbidity 1996 - 2013, by gender, age group and race/ethnicity, CDC WONDER Online Database. The database provides a set of variables to draw down the information in a report format.

Due to the possibility of changing information, such as cases being reported much later than the occurance, it is prudent to provide the date and time of access. The CDC WONDER Online Database Query was conduced on Oct 17, 2015 at 8:30:27 AM.

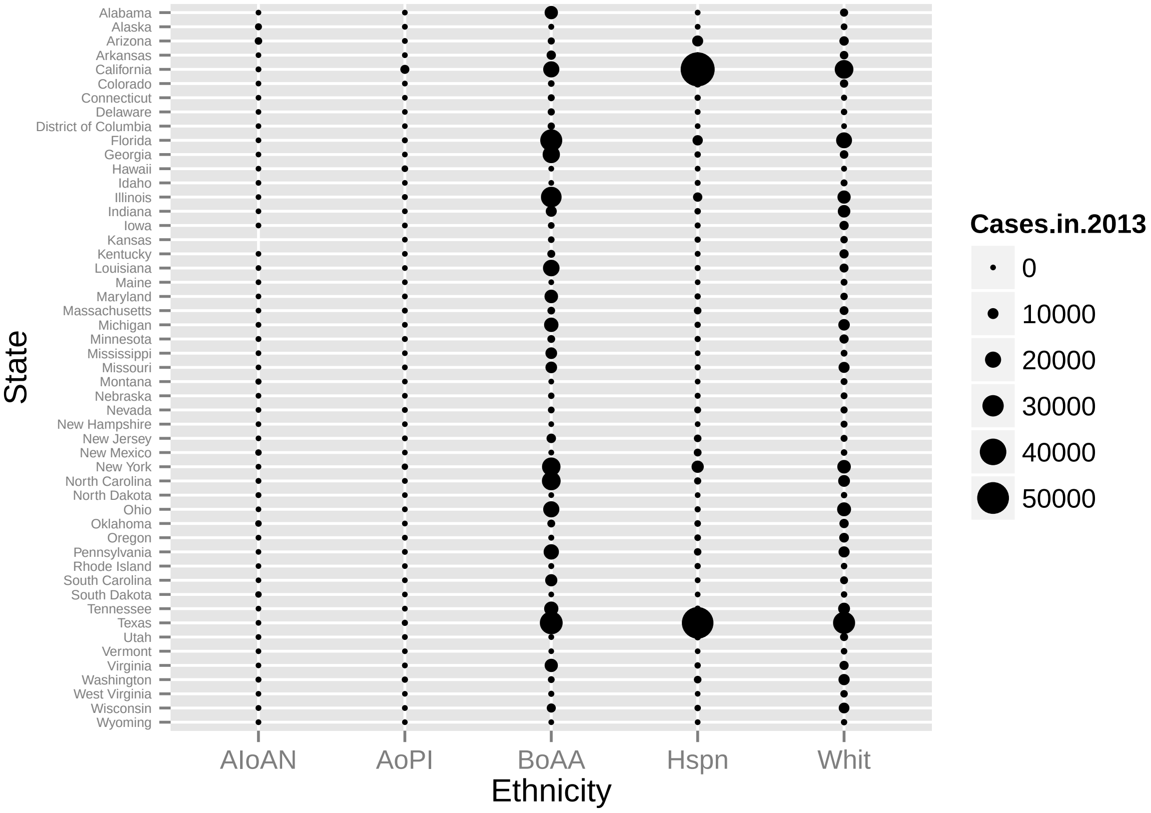
Part 3 - Exploratory data analysis:

As we discovered earlier the difference between actual cases becomes quite large because of the variance of the populations as seen in the below graph.

```
cases_table_visual <- aggregate(list('Cases in 2013' = df$STD.Cases),
                by = list(Ethnicity = df$Race.Ethnicity,'State' = df$State), FUN=sum)
cases_table_visual <- cases_table_visual[order(cases_table_visual$Ethnicity),]
```

```
par(mfrow=c(1,1))
cases_table_visual <- within(cases_table_visual,
                             State <- ordered(State, levels = rev(sort(unique(State)))))

ggplot(cases_table_visual, aes(Ethnicity,State)) +
  geom_point(aes(size = Cases.in.2013)) +
  scale_x_discrete(label=abbreviate) +
  theme(axis.text.y = element_text(size = 5))
```

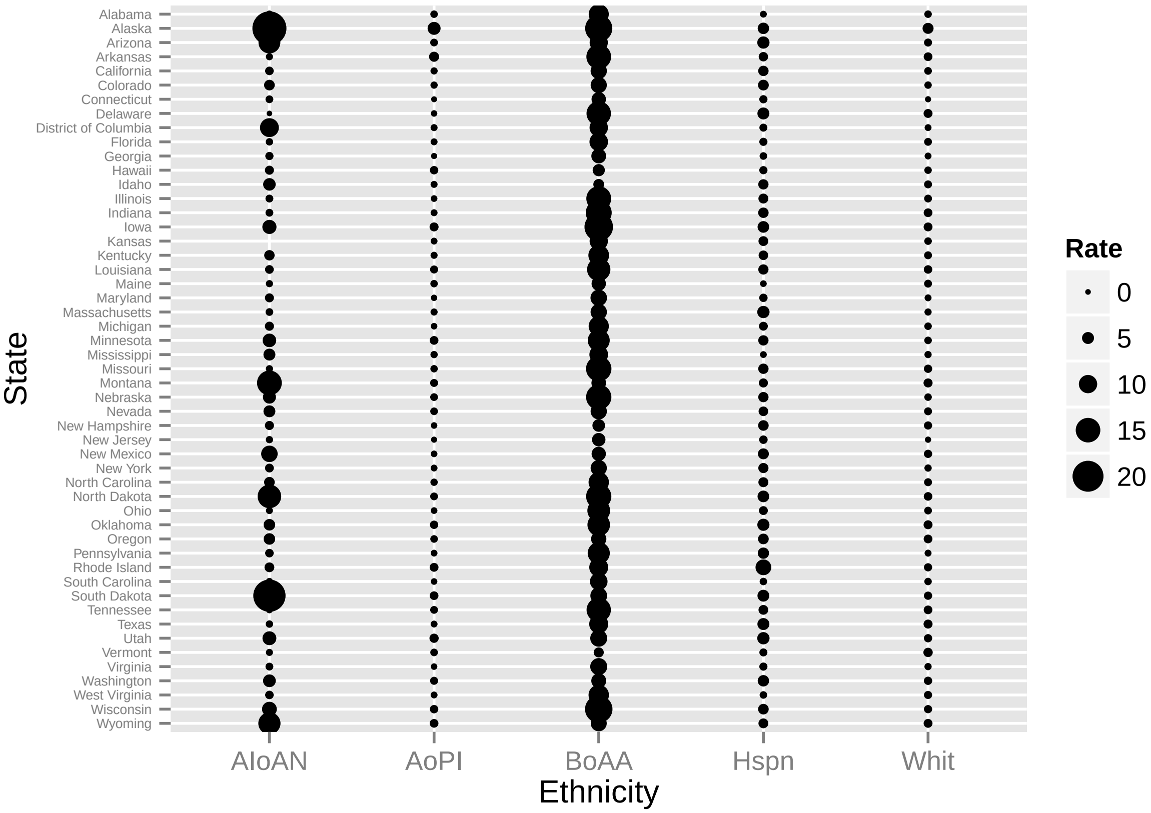


In order to adjust for the population variance among groups we will look at the rates of cases as determined by the ethnicity and state allowing us to compare across groups.

The `rates` in our data set are for every 1,000 people. That is, if a rate for a group was 2 then 2 out of 1,000 for that group's population would have contracted the infection. As we can see the rates of infection provides a much different picture when we control for total population variance.

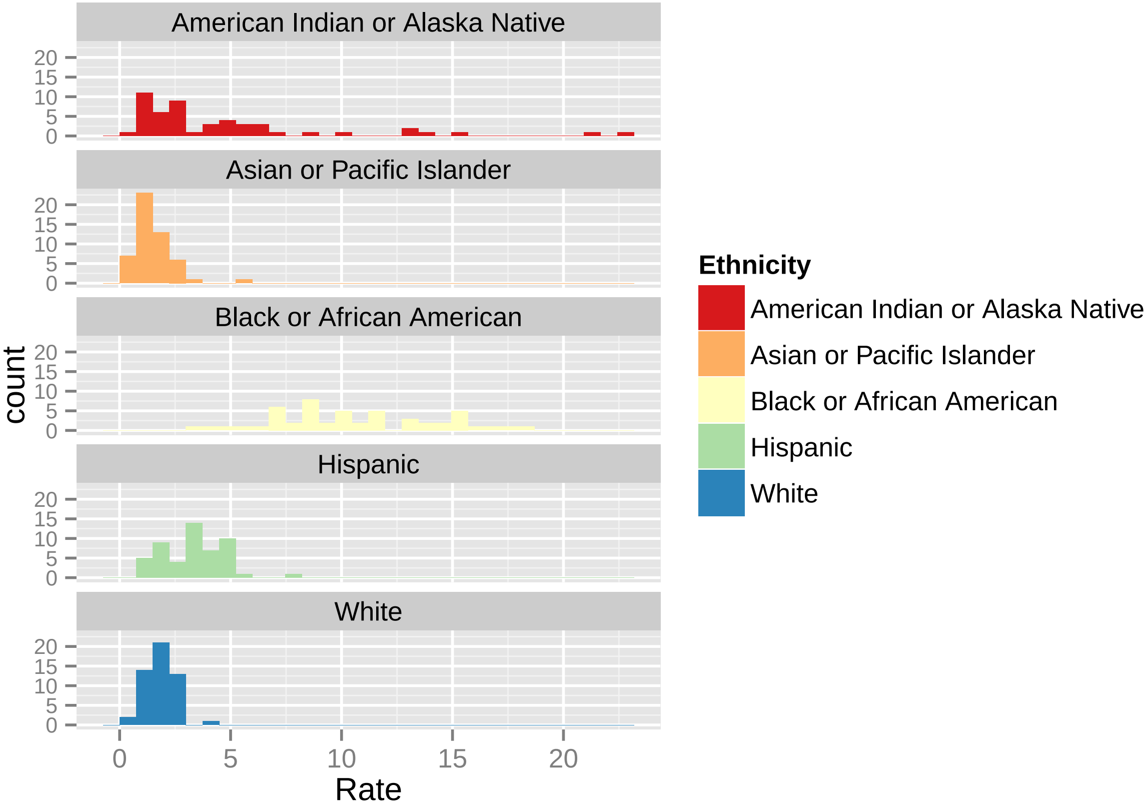
```r
rates_table <- aggregate(list('Cases' = as.numeric(df$STD.Cases),
                              "Population" = as.numeric(df$Population)),
                                by = list(Ethnicity = df$Race.Ethnicity,
                                          State = df$State), FUN=sum)
rates_table$Rate <- (rates_table$Cases/rates_table$Population) * 1000
rates_table <- rates_table[order(rates_table$Ethnicity),]
rates_table <- within(rates_table,
                      State <- ordered(State, levels
                                       = rev(sort(unique(State)))))
```

```r
par(mfrow=c(1,1))
ggplot(rates_table, aes(Ethnicity, State)) +
  geom_point(aes(size = Rate)) +
  scale_x_discrete(label=abbreviate) +
  theme(axis.text.y = element_text(size = 5))
```



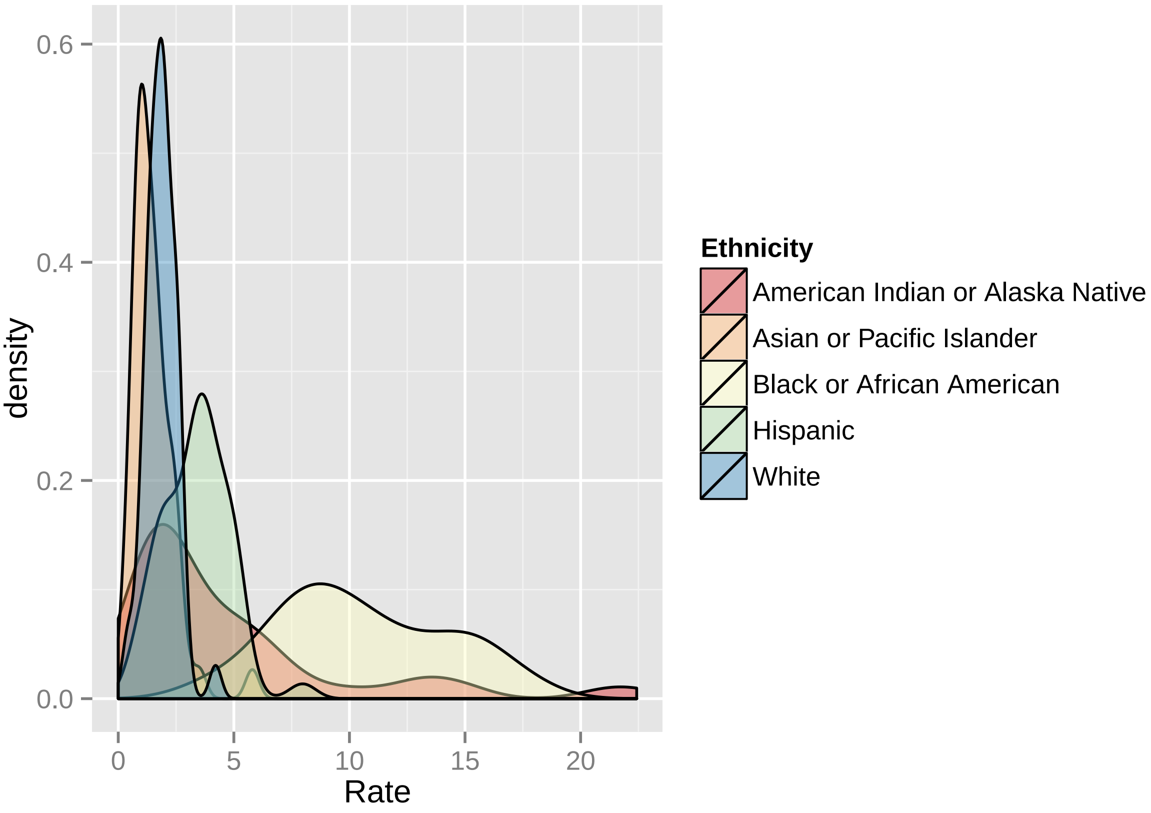
The below histogram provides evidence that the difference of infection rates is significantly different among the populations.

```
y <- ggplot(rates_table, aes(Rate, fill = Ethnicity))
y + geom_histogram() + facet_wrap(~Ethnicity, ncol = 1) +
  scale_fill_brewer(palette="Spectral") +
   theme(axis.text.y = element_text(size = 8))
```



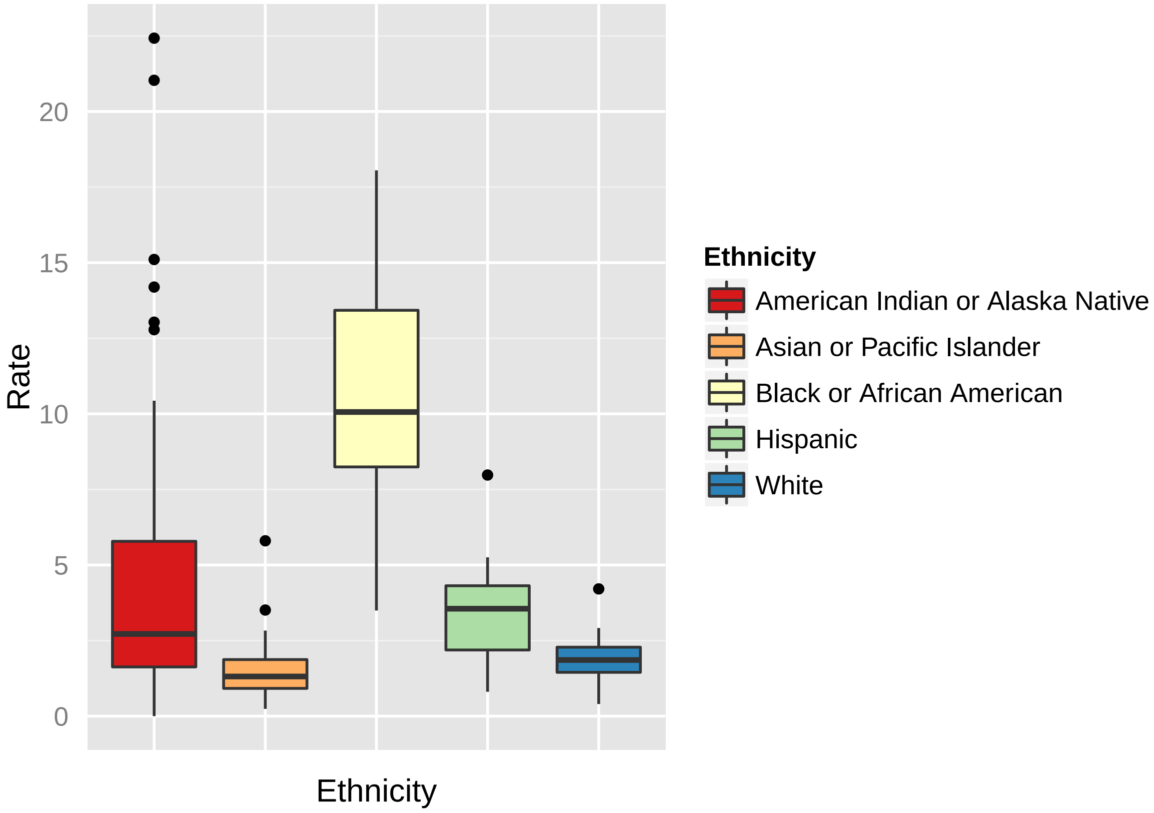
The below density plot provides insight to the differences in the rate observations by each ethinic group. By overlaying the density plots we can more clearly see the differences in rates among the population.

```
x <- ggplot(rates_table, aes(Rate, fill = Ethnicity))
x + geom_density(alpha = 0.4) + scale_fill_brewer(palette="Spectral")
```



The boxplot graph of the data set is fairly telling that the rates of infection is not similar among the ethnicity groups.

```
x <- ggplot(rates_table, aes(Ethnicity, Rate, fill = Ethnicity))
x + geom_boxplot() + scale_fill_brewer(palette="Spectral") +
     theme(axis.ticks = element_blank(), axis.text.x = element_blank())
```



The below is the summary rates for ethnicity by state:

	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
American Indian or Alaska Native	50	4.83	5.02	2.72	3.79	2.26	0.00	22.43	22.43	1.89	3.18	0.71
Asian or Pacific Islander	51	1.50	0.93	1.31	1.39	0.74	0.24	5.80	5.56	2.08	6.85	0.13
Black or African American	51	10.59	3.64	10.06	10.53	4.18	3.49	18.05	14.56	0.22	-0.91	0.51
Hispanic	51	3.38	1.40	3.55	3.37	1.49	0.81	7.98	7.17	0.37	0.56	0.20
White	51	1.86	0.66	1.86	1.86	0.63	0.40	4.21	3.81	0.53	1.62	0.09

The below is the summary rates for ethnicity by region. I have collapsed the states into regions (Northeast, Midwest, South, West) in order to conclude if there appears to be a difference in the average rate between larger regions. Each state has such unique difference I wanted further conclusive evidence that even adjusting for states we would see a difference in the average rate:

	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
American Indian or Alaska Native	4	5.71	3.52	5.61	5.71	4.26	2.15	9.45	7.30	0.03	-2.33	1.76
Asian or Pacific Islander	4	1.12	0.28	1.11	1.12	0.32	0.81	1.44	0.63	0.03	-2.14	0.14
Black or African American	4	10.47	2.59	9.68	10.47	1.54	8.40	14.12	5.72	0.51	-1.88	1.29
Hispanic	4	3.62	0.24	3.58	3.62	0.22	3.39	3.93	0.55	0.28	-2.05	0.12
White	4	1.68	0.37	1.86	1.68	0.03	1.13	1.88	0.75	-0.75	-1.69	0.18

Even as we adjust for larger regions we can clearly see a difference in the mean rates amoung our groups.

Scope of Inference

The population of interest is individuals that tested positive for Chlamydia. The study is limited to people that underwent testing and aggregated to the state level so the findings could be compared to only those that are tested for Chlamydia across US regions. A significant limitation is that we do not know how many people were tested and also we do not know how many individuals have the infection but are asymptomatic.

Unfortunately, We will not be able to infer causality because Chlamydia is an asymptomic infection and the sample that is identified is a bias sample. The sample is limited to only those individuals that received testing and further research would needed to look at factors associated with being tested, such as age and gender.

Part 4 - Inference:

Hypothesis

$$H_o : \mu_{American\ Indian\ or\ Alaska\ Native} = \mu_{Asian\ or\ Pacific\ Islander} = \mu_{Black\ or\ African\ American} = \mu_{Hispanic} = \mu_{White}$$

$$H_A : The\ average\ rate\ varies\ across\ some\ (or\ all)\ groups$$

The confidence interval for the rates among the ethnicities by the states is below:

```r
confidence_interval <- function(ethnicity){
  x <- subset(sum_stat_m, rownames(sum_stat_m) == ethnicity)
  mean <- x$mean
  standard_de <- x$sd
  sample_size <- x$n
  error <- qnorm(0.975)*standard_de/sqrt(sample_size)
  left <- mean - error
  right <- mean + error
  paste("Confidence interval for the rate of infection for ",
        ethnicity," is (",round(left,4)," , ",round(right,4),")", sep="")
}
```

```r
confidence_interval("American Indian or Alaska Native")
```

```
[1] “Confidence interval for the rate of infection for American Indian or Alaska Native is (3.4386 , 6.2214)”
```

```r
confidence_interval("Asian or Pacific Islander")
```

```
[1] “Confidence interval for the rate of infection for Asian or Pacific Islander is (1.2448 , 1.7552)”
```

```r
confidence_interval("Black or African American")
```

```
[1] “Confidence interval for the rate of infection for Black or African American is (9.591 , 11.589)”
```

```r
confidence_interval("Hispanic")
```

```
[1] “Confidence interval for the rate of infection for Hispanic is (2.9958 , 3.7642)”
```

```r
confidence_interval("White")
```

```
[1] “Confidence interval for the rate of infection for White is (1.6789 , 2.0411)”
```

As we can see from the above confidence intervals, several ethnicities have no overlap. Using 95% confidence intervals we can conclude that there is a difference in the means of the rates of incidence. We can therefore reject the null hypothesis that the means among groups is the same.

We can futher review difference in means through an ANOVA test.

```
Aov_Ethnicity <- aov(Rate ~ Ethnicity, data=rates_table)
summary(Aov_Ethnicity)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)    
## Ethnicity     4   2771   692.6   83.66 <2e-16 ***
## Residuals   249   2062     8.3                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA test for just Ethnicity , we can see that by ethnicity grouping we have a extremely low p value at 0.0000000000000002. This is essentially zero and at zero the p value is less than the .05 significance level so we can reject the null hypothesis.

```
Aov_State <- aov( Rate ~ State, data=rates_table)
summary(Aov_State)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## State        50    662   13.23   0.644  0.967
## Residuals   203   4171   20.55               
```

The ANOVA test for just State is interesting as we would be unable to reject the null hypothesis on state rates alone due to the large p value at `0.967`. It seems that the differences in rates is not as well explained by geography alone.

```
Aov <- aov(Rate ~ Ethnicity + State , data=rates_table)
summary(Aov)
```

```
##              Df Sum Sq Mean Sq F value  Pr(>F)    
## Ethnicity     4 2770.6   692.6  98.424 < 2e-16 ***
## State        50  661.2    13.2   1.879 0.00124 ** 
## Residuals   199 1400.4     7.0                    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA test including both ethnicity and geography provides a through analysis of the data set and provides the information to reject the null hypothesis and conclude there are differences between ethnicity and infection rates in the US.

While the ANOVA information appears promising we must check for the conditions of inference. The sample is not 10% of the population as we are gathering the entire population data for those that test positive for the infection in the US. However, due to the asymptomic nature of the infection, many who are infected will not be treated so we cannot be sure of the true population of those current infected with Chlymdia. Also, we have to consider the possiblity that we are capturing many pairs of partners who have give each other the infection resulting in possible bias. As we discovered in our data analysis the data set does appear to be near normal for the infection rates across states and the variance is nearly constant for all groups.

Using simulation to observe inference of infection for populations. In this simulation we take random samples from our rates data set to see the impact in the ANOVA results. The expectation is that if the null hypothesis is true we would not expect to see a significant difference in the mean rates between ethnicities:

```r
sample_1 <- rates_table[sample(nrow(rates_table), 50), ]
fit_sim_1 <- aov( Rate ~ Ethnicity + State, data= sample_1)
summary(fit_sim_1)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)    
## Ethnicity    4  410.3  102.57  37.562 1.2e-07 ***
## State       30   96.8    3.23   1.182   0.376    
## Residuals   15   41.0    2.73                    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
sample_2 <- rates_table[sample(nrow(rates_table), 75), ]
fit_sim_2 <- aov( Rate ~ Ethnicity + State, data= sample_2)
summary(fit_sim_2)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)    
## Ethnicity    4  984.4  246.10  50.535 1.16e-12 ***
## State       41  187.5    4.57   0.939     0.58    
## Residuals   29  141.2    4.87                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Our results remain unchanged, the state variation doesn't provide significant evidence of differences but ethnicities does provide significant evidence.

Part 5 - Conclusion:

We have found that variation does exist between ethnic groups in the US. However, the data used for this data project has a number of data limitations and was not as appropriate for this type of analysis as I had hoped. Unfortunately, the data set has the summary statistic population of those tested positive for Chlamydia which does not allow for the more typical data analysis of reviewing each case and make inferences of the general population. I was unable to evaluate negative test results and positive test results which would have allowed for a more thorough analysis. In general, I learned how critical it is to review a data set for its inference capabilities before conducting a lenthly analysis. In the future, I believe that a data set of actual observations would prove more useful than a data set of summary information observations.

As we can see in the data analysis section there is a clear difference in the rate of infection for Black or African American than other groups. Also, our confidence intervals for ethnicity has several ehtnicities with intervals that do not overlap. This helps us understand that the true population mean is statistically different from the other means. The greatest difference was seen between Asian or Pacific Islander and Black or African American. This is an interesting observation to explore as there could be a number of factor contributing to this obviously large difference. Such as, why do we see such high Chlamydia infections in this ethnic group compared to the other? Is this group simply more aware of the asytompatic nature of the condition and choose to be tested for it more often, or could this group communicate their infections to their partners more effectively?

References:

[1] (n.d.). Retrieved November 18, 2015, from http://www.cdc.gov/std/chlamydia/chlamydia-factsheet-june-2014.pdf

[2] Chesson, Harrell W. et al.. “The Estimated Direct Medical Cost of Sexually Transmitted Diseases Among American Youth, 2000”. Perspectives on Sexual and Reproductive Health 36.1 (2004): 11-19.

[3] “CDC Grand Rounds: Chlamydia Prevention: Challenges and Strategies for Reducing Disease Burden and Sequelae”. “CDC Grand Rounds: Chlamydia Prevention: Challenges and Strategies for Reducing Disease Burden and Sequelae”. Morbidity and Mortality Weekly Report 60.12 (2011): 370-373. Web. . .